Demonstration of online auditory scene analysis

J.D. Krijnders T.C. Andringa

Auditory cognition group, Rijksuniversiteit Groningen, P.O.Box 407, 9700 AK Groningen

Abstract

We show an online system for auditory scene analysis. Auditory scene analysis is the analysis of complex sounds as they occur in natural settings. The analysis is based on models of human auditory processing. Our system includes models of the human ear, analysis in tones and pulses and grouping algorithms. This systems forms the basis for several sound recognition tasks, such as aggression detection and vowel recognition.

Current sound recognition systems function well in controlled spaces with one, well defined source. But as soon as multiple sound sources are present the performance drops rapidly[3]. Humans on the other hand seems to have little problems recognising sounds in the presence of noise[5] or other sources. One of the goals of (computational) auditory scene analysis[7] is to try to bridge this gap by basing its methods on what is known about human auditory processing. We will demonstrate our system based on techniques from auditory scene analysis augmented by algorithms that select spectro-temporal area's with positive signal-to-noise ratio.

1 Methods

The model demonstrated is based on a transmission line model of the human ear[1]. Its latency is lower than filterbank implementations[2], which makes it more suitable for online models. The output is squared, leaky-integrated and down-sampled which results in a energy representation. This representation is called a cochleogram.

To find areas of the cochleogram where tones or pulses dominate, we filter it with two segment-dependent filters. These filters match the shape of the tone response and the pulse response of the cochleogram up to two standard deviations of white noise under the peak of the response. The result is two representations, one which indicates the pulsality, the other indicating the tonality of each time-frequency combination of the signal. These measures can be interpreted as local signal-to-noise ratios under the assumption that the signal is a pulse, resp. a tone.

In both representations the neighboring local maxima are joined to form regions of the spectrum with are likely to be one continuous tone or pulse. These regions we call signal components. Tonal signal components are then grouped into harmonic complexes, this can increase the robustness of the signal components and leads to groups of signal components that are highly likely to stem from a single source

2 Discussion

The cochlea model is successfully applied in several sound monitoring projects in the Netherlands[6]. The complete system is used in several projects ranging from aggression detection[8] to environmental sounds recognition[4].

3 Demo requirements

The demo runs continuously, but with explanation will take about ten minutes to visit. Requirements would include a large screen or beamer and a table for it.

References

[1] H. Duifhuis, H.W. Hoogstraten, S.M. Netten, R.J. Diependaal, and W. Bialek. *Cochlear Mechanisms: Structure, Function and Models*, chapter Modelling the cochlear partition with coupled Van Der Pol oscilators, pages 395–404. Plenum, New York, 1985.

[2] Toshio Irino and Roy Patterson. A time-domain, level-dependent auditory filter: The gammachirp. *Journal of the Acoustical Society of America*, 101(1):412–419, January 1997.

[3] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, July 1997.

[4] M.E. Niessen, L. Van Maanen, and T.C. Andringa. Disambiguating sounds through context. In *Proc. IEEE International Conference on Semantic Computing*, 2008.

[5] H.J.M. Steeneken. *On measuring and predicting speech intelligibility*. PhD thesis, University of Amsterdam, 1992.

[6] P.W.J. van Hengel and T.C. Andringa. Verbal aggression detection in complex social environments. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 15–20. IEEE, 2007.

[7] DeLiang Wang and Guy J. Brown. *Computational Auditory Scene Analysis*. John Wiley and Sons, Holoken, NJ, 2006.

[8] W. Zajdel, J.D. Krijnders, T.C. Andringa, and D.M. Gavrila. Cassandra: audio-video sensor fusion for aggression detection. In *Proc. IEEE Conference on Advanced Video and Signal Based Surveillance AVSS 2007*, pages 200–205, 2007.